Luonnollisen kielen tilastollinen käsittely

T-61.281 (3 ov) L

Kevät 2003

Luennot: **Timo Honkela**
Laskuharjoitukset: **Vesa Siivola**

Luentokalvot: Krista Lagus ja Timo Honkela

1. Yhteenveto

- Pikakertaus
- Tuntematon sana kontekstissa - tulokset
- Väritehtävä
- Harjoitustyöstä

1.1 Pikakertaus

- Yleistä kielitieteestä ja tilastollisesta lähestymistavasta
  - Kategoriset versus jatkuvat representaatiot
  - Probabilistiset esitystavat
  - Datasta oppiminen

- Matemaattisia perusteita
  - Ehdollinen todennäköisyys
  - Riippumattomuus
  - Bayesin kaava, bayesläinen päätösteoria
  - Jakaumat
  - Informaatioteoria
  - Entropia
  - MDL - minimum description length

- Yleisen kielitieteen perustietoja
  - Kielen tasot ja alueet
  - Kielioppiformalismit
  - Korpustyöskentely
  - Tekstin esikäsittely
  - Morfologia-analyysi
  - Monitulkintaisuus ja disambiguointi
  - Taggaus

- Kollokaatiot
  - Frekvenssitiedot, sanaluokkasuodatus
  - Hypoteesien testaus: T-testi, khii-toiseen, uskottavuuksien suhde
  - Pisteittäinen yhteisinformaatio

- Tiedonhaku
  - Täsmäosumat, rankkaus
  - Rakenne: käänteisindeksi, sulkulista
  - Mittoja: saanti ja tarkkuus
  - TREC
  - Vektoriavaruusmalli
  - Termien painotus (esim. tf.idf)
  - Latenttien muuttujien menetelmät, LSI
  - Riippumattomien komponenttien analyysi (ICA)
  - Dimension pienennys

- N-grammimallit
  - Ekvivalenssiluokat
  - Estimointi, esim. Good-Turing
  - Tasoitusmenetelmät (nähdyt vs ennen näkemättömät)
  - Estimaattorien yhdistäminen
  - Mallien estimoinnista yleisesti: vertailu, opetusjoukko-testijoukko, ristiinvalidointi

- Markov-mallit
  - Näkyvät Markov-mallit: äärellinen horisontti, stationaarisuus
  - HMM: tilajono ja havaintojono
  - Havaintojonon tuottaminen
  - Annetun havaintojonon todennäköisyyden laskeminen
    - Forward-algoritmi (eteenpäinlaskenta)
    - Backward-algoritmi (taaksepäinlaskenta)
  - Todennäköisimmän tilajonon etsiminen
    - Viterbi-algoritmi
  - HMM:n parametrien estimointi
  - Soveltaminen, mm. puheentunnistus

- Sanaluokkien taggaus
  - Markov-taggerit
  - Tuntemattomien sanojen käsittely
  - HMM-taggerit
  - Muunnoksiin perustuva taggaus

- Probabilistinen jäsentäminen ja PCFG:t
  - Kontekstivapaat kieliopit (CFG)
  - Chomskyn hierarkia kielille
  - Lauseen todennäköisyyden laskeminen
    - Inside-algoritmi
  - Todennäköisimmän jäsennyksen valinta lauseelle
  - PCFG:n parametrien estimointi joko jäsennetystä (ohjattu oppiminen) tai jäsentämättömästä datasta (ohjaamaton oppiminen)
  - Probabilistinen leksikalisoitu CFG

- Leksikaalinen semanttinen tieto ja semanttinen samankaltaisuus
  - Wordnet-tietokanta
  - Temaattiset roolit
  - Verbien argumentinvalintapreferenssit (selectional preferences)
  - Semanttinen samankaltaisuus
  - Menetelmiä: KNN, vektorietäisyydet, yhteisesiintymätiedot

- Sananmerkitysten yksikäsitteistäminen
  - Eri oppimisperiaatteet
  - Menetelmien onnistumisen mittaaminen
  - Piirteiden valinta
  - Ohjattu disambiguointi
    - Bayesläinen luokitin
    - Informaatioteoreettinen lähestymistapa
    - Sanakirjapohjainen disambiguointi
    - Kaksikielisen aineiston käännöksiä hyödyntävä menetelmä
  - Ohjaamaton merkitysten ryhmittely
    - EM-algoritmi disambiguoinnissa
    - klusterointi

- Klusterointi
  - Parametriset (EM, mikstuurimallit) vs. epäparametriset menetelmät
  - Hierarkkiset (esim. single-link clustering) vs. litteät (esim. K-means) menetelmät
  - Kovan vs. pehmeän klusteroinnin tekevät menetelmät
  - Klusterointiongelman ratkaisun vaiheet

- Tekstin luokittelu
  - Naive Bayes -luokitin
  - Päätöspuut
  - Monikerrosverkko

- Tilastollinen konekääntäminen
  - Perinteisesti: muunnosmenetelmät
  - Tekstin linjaaminen
  - Sanakirjan indusointi
  - Käännös

1.2 Tuntematon sana kontekstissa

- 'Rauma kaupunki rakensi oman X:n' - teknologiakeskus, satama, pomppulinna, museo, tietoverkko, ydinvoimala, konserttitalo, lumilinna, stadion, jäähalli, museo, musiikkitalo, jäähalli, uimahalli, lumilinna
- 'Rauma kaupunki rakensi oman X:n 1897' - kaupungintalo, satama, kaupungintalo, kaupungintalo, hiilivoimala, satama, oikeustalo, kauppahalli, palolaitos, kaupungintalo, kaupungintalo, puhelinkeskus, satama, -, linna
- 'Agricolan kirja-arvostelut X ONNEEN' - kirjasto, tie, tie, tie, tie, tie, huomenna, tie, kirjasto, oikotie, museo, oikotie, tie, tie, tie
- rautatiemuseo, rautatie, tie, tie, rautatie, rautatie, rata, rata, rata, oikotie, rautatie, rautatie, rata, rata, rata

1.3 Väritehtävä

- tehdään pareittain
- henkilö 'A' kirjoittaa, millä sanalla tai fraasilla hän kuvaisi näkemänsä värin ja kertoo tämän henkilölle 'B' näyttämättä väriä missään vaiheessa toiselle
- henkilö 'B' valitsee värikartalta värin, joka hänen mielestään vastaa parhaiten henkilön 'A' määrittelyä ja kirjoittaa ko. kohtaan järjestysnumeron

1.4 Harjoitustyöstä

- kysymys-vastaus -järjestelmä
- probabilistinen jäsennys
- sanojen merkitysten yksikäsitteistäminen
- tiedonhaku